Combining Twitter and Earth Observation Data for Local Poverty Mapping

Lukas Kondmann, Matthias Haeberle and Xiao Xiang Zhu
Technical University of Munich (TUM)
& German Aerospace Center (DLR)
{lukas.kondmann,matthias.haeberle,xiaoxiang.zhu}@dlr.de

Abstract

Accurate and timely data on economic development is essential for policy-makers in low- and middle-income countries where such data is often unavailable. To fill this gap, existing approaches have used alternative data sources to proxy for levels of local development such as satellite imagery or mobile phone data. In this paper, we underline the power of an underrated data source for poverty mapping: Geolocated tweets. We show that the number of tweets in a region as singular input can already explain 55 % of the variation in local wealth in Sub-Saharan Africa with a simple Random Forest model. When nighttime light and Twitter usage information are combined as inputs to a Random Forest model they already explain 65% of the variation in local wealth which is in the range of state-of-the-art neural network architectures based on satellite images. Our results show that the naive combination of these data sources in a random forest is already competitive in performance and more elaborate fusion approaches are a promising direction to advance the accuracy of poverty mapping.

1 Introduction

To this day, more than 700 million people still live in extreme poverty.[1] Although ending world poverty is the first of the 17 UN Sustainable Development Goals it is difficult for policy-makers in practice to monitor the progress towards this goal. This is because data about local poverty is often missing, unreliable, or outdated in developing countries which makes it hard to design and target policy interventions effectively [8].

One prominent example of this is satellite data where mostly nightlight images (NTL) [3, 4, 6, 18] but also daytime images [7, 9, 12, 19] have served as input for poverty mapping. Beyond earth observation data, a variety of inputs measured on the ground have been used to estimate local development. Blumenstock et al. [1] pioneer the use of mobile phone metadata to predict local wealth. Building on this work, subsequent contributions have combined mobile phone data with satellite images for poverty mapping in Rwanda [10], Senegal [13] and Bangladesh [15]. Other inputs that have shown promise for poverty mapping include geolocated Wikipedia articles [14], connectivity and device information from Facebook's advertising platform [5], and open street map data [17, 20].

Another rich data source of human activity is social media and specifically Twitter data. Tweets are known to be a helpful information source for the related task of population mapping [11], yet they are surprisingly underutilized for poverty mapping. To close this gap, we analyze the possibility to predict local development indicators based on geolocated tweets in this paper. Specifically, we test if tweet counts and text information can be used to predict local poverty levels in sub-Saharan

[1] https://www.un.org/sustainabledevelopment/poverty/

ML4D Workshop at 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

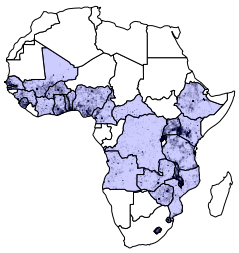

Figure 1: DHS Cluster locations in our dataset. Included countries are displayed in blue.

Africa with Random Forests. For this purpose, we use a dataset of average local wealth across 20,000 villages by Yeh et al. [19] and combine it with information on Twitter activity and nightlight intensity in these villages. Our contribution is two-fold:

1. We show that twitter information is a promising source of information for mapping local economic development. Specifically, the number of tweets on its own is already quite predictive of local wealth levels with an R^2 of 55% in a single-input Random Forest model.
2. Combining tweet and nightlight information in a Random Forest model explains up to 65% (R^2) of the variance in local poverty which is already in the range of state-of-the-art convolutional architectures (68%) based on NTLs with a computationally cheap and easy-to-implement approach.

These results serve as an exploratory study to underline the potential in combining EO and social media data with machine learning for poverty mapping and we aim to investigate the possibilities of such approaches in more depth in future work.

2 Data and Methods

We assemble a unique dataset of local wealth levels, twitter activity, and nighttime lights in sub-Saharan Africa for our analysis. The data on average local wealth is taken from Yeh et al. [19] who use survey data about household asset ownership from the Demographic and Health Surveys (DHS) as wealth information. Households are surveyed in clusters of about 10km $\times$ 10km in size which is roughly equivalent to the size of a village. Reported wealth levels are the average of the surveyed households within a cluster for about 20,000 villages across 23 African countries between the years 2009 and 2016. Figure 1 displays the geographical distribution of the villages across the African continent. The data covers large parts of sub-Saharan Africa from Mali in the north-west to Lesotho in the south.

Nighttime light intensity values are taken from the Defense Meteorological Satellite Program/Operational Linescan System (DMSP-OLS) annual stable lights V4 composite of 2013 by the National Oceanic and Atmospheric Administration (NOAA). DMSP-OLS NTLs measure nightlight intensity in a range from 0 to 63 in cells of 30 arc seconds which is equivalent to about 1km $\times$ 1km. While it would in theory be preferable to collect NTL data from the same year as the survey data, DMSP-OLS data is only collected until 2013. Even before 2013, there is subtle differences in the satellite used and the overpass time which makes DMSP data hardly comparable across years. Hence, we restrict our usage of NTLs to the year 2013 only. As one cluster spans about 10km $\times$ 10km, we record the average NTL intensity of about 100 cells in the cluster for our analysis.

Geolocated tweets are collected from the random sample stream of the official Twitter API from November 2018 to February 2020. In total, our sample contains over 4 million tweets across Africa. Even though twitter users can not be seen as representative of the village population, we want to analyze if this data source can nevertheless be useful for this task. We average the number of tweets in each cluster.

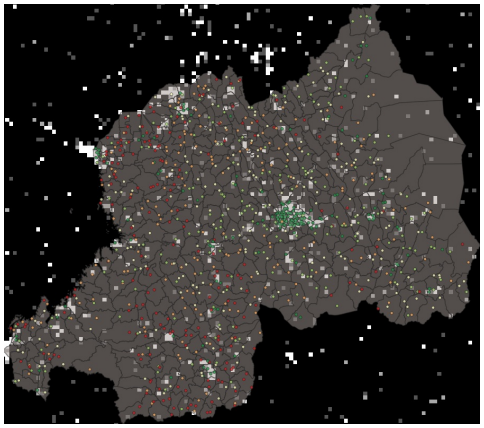

Figure 2: Tweet counts (black and white) & Rwandan clusters (quintiles of wealth from red to green)

Further, we also extract a 300-dimensional fastText [2] embedding of the English language tweets in our sample and average them per cluster to get the average 300-dimensional feature vector representing the contents of the tweets in the cluster. This gives us a dataset of about 20,000 villages with average wealth levels of households, average NTL intensity in the region, tweet summary statistics, and the average content of the English tweets in the cluster. Figure 2 visualizes the key variables of our dataset for Rwanda. Tweet density is the bottom layer from black to white which is overlayed by a map of Rwandan counties. On top are DHS clusters in our dataset which are colored by wealth quintile from red to green. Figure 2 underlines the spatial relationship between tweets and wealth we are exploiting: A high density of tweets (bright pixels) corresponds to a comparably high wealth level (green clusters). This is especially visible in and around Kigali in the center of the map but seems to be a general pattern.

We can use this spatial relation and predict local poverty from Tweets. For this, we use a Random Forest Regressor (RF) even though more complex, neural network-based algorithms would potentially be even more promising for this task. We nevertheless choose a RF model to demonstrate the feasibility of this approach even without computationally expensive training and the necessity of large labeled datasets. As clusters in the dataset can be geographically close or in extreme cases even overlapping, we follow Yeh et al. [19] in their spatial split of the clusters in five geographically exclusive folds for each country to ensure no label leakage between the folds. We do cross-validation over these folds where three folds serve as training set, one as validation set for tuning hyperparameters and one as final test set.

We test several versions of our model with different input data. First, we separately test only tweet counts, only NTL intensity, and only Tweet text features as input. Then, we combine these features together and use them as combined input to the model. To be consistent with Yeh et al. [19], all models are evaluated based on R^2 averaged over the respective test folds. We tune the hyperparameters with a random search over the following specifications: Maximum depth (2 to 30), number of estimators (10 to 50), and maximum features (2 to 100). We compare our results with CNN NTL which is one of the models from Yeh et al. [19] using upsampled NTL images of two different sensors as input. Further, CNN transfer is based on Jean et al. [7] who use NTLs as an indicator for economic development. This transfer learning step is necessary because of a lack of data on wealth which prevents training end-to-end.

3 Results

Table 1 presents the average R^2 values on left-out test folds of our models and compares them with existing results. At first, using only the counts of tweets within a cluster already explains 55 % of the wealth variation across the 20,000 villages. This ratio cannot be improved by using more elaborate text features which implies that the RF model may not be able to make use of the text embeddings beyond the fact how much people tweet. Using tweet counts is even in proximity to, yet does not surpass, the RF model based on NTLs which explains 62% in wealth variation. NTLs

Table 1: Share of explained variance in local wealth by model

Model Name	Average R^2	Authors
RF Tweet Counts	0.55	Ours
RF Tweet Text	0.55	Ours
RF NTL	0.62	Ours
RF NTL + Tweet Count	0.64	Ours
RF NTL + Tweet Text	**0.65**	Ours
RF NTL + Tweet Text & Counts	**0.65**	Ours
CNN NTL	0.68	Yeh et al. [19]
CNN Transfer	0.59	Jean et al. [7]

are a highly preprocessed information source and widely established as a good predictor of local development [18, 3]. The fact that simple tweet counts already reach a predictive power close to NTLs is encouraging. The intuition for this result is already visible in Figure 2: Tweets and local development are closely related and this spatial pattern can be exploited for modeling local wealth. The number of tweets might serve not only as a proxy for population density but could also be related to socio-economic status as this is likely to influence tweeting behavior [16].

The combination of tweets and NTLs leads to improved results over the single input models with 64% (Tweet counts only) and 65% (Text features) in explained variation of local wealth which is 2-3 percentage points higher than the single NTLs model. Again, it seems like our model can not make use of additional information in the language features beyond Tweet quantities as adding both does not improve R^2 beyond 65%. This value can be seen as a lower bound of what is achievable with the combination of these data sources since we do not tap the potential of large, pre-trained neural networks to make sense of the large body of image and text data here. Instead, we provide a fast, computationally cheap, and easy-to-implement method to use this data for poverty mapping. This approach is already quite competitive compared to CNN based approaches without the necessity of large scale training. We surpass the performance of the pioneering transfer learning strategy which has been shown to be superior to mobile-phone data approaches by Jean et al. [7] with a margin of 6 percentage points. Further, we are slightly below but in the range of the current state-of-the-art CNNs based on two different NTL sensors (68%).

These results underline the vast potential in twitter data to map local levels of poverty that might be especially beneficial in combination with NTL sensors. In combination, the two data sources achieve an R^2 of 65% for wealth mapping in sub-Saharan Africa. Although obtained on different datasets, this also exceeds 63% R^2 obtained with the combination of open street data and NTLs in the Philippines [17] even though our model is fitted over 23 countries. Hence, twitter data might be at least as promising as OSM data for the combination with NTLs. In future work, we aim to design a more advanced fusion architecture for tweets and NTLs which makes use of the full richness of information given in image and text data from EO and Twitter.

4 Conclusion

In this paper, we present the potential of twitter activity as input for poverty mapping in a large dataset of local wealth in 20,000 villages across sub-Saharan Africa. For these villages, we collect a small random sample of tweets in Africa from November 2018 to February 2020 from the Twitter API, DMSP-OLS nightlight intensity values from 2013, and average local wealth levels based on household surveys. We predict local wealth with a Random Forest regressor based on the number of tweets in our sample only and find that this can already explain 55% in the variation in local wealth which demonstrates the richness of information in tweet behavior. Combining nightlights and twitter data explains up to 65% percent of local variation which is a notable increase compared to using nightlights only. Since an R^2 of 65% is in the range of current state-of-the-art deep learning models based on multi-source nightlight data inputs (68%), we conclude that the fusion of social media and earth observation data has great potential to further boost poverty mapping efforts. In future work, we plan to investigate deep learning-based fusion architectures that are able to combine the power of text and image data for this task in more depth.

References

[1] J. Blumenstock, G. Cadamuro, and R. On. Predicting poverty and wealth from mobile phone metadata. *Science*, 350(6264):1073–1076, 2015.

[2] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

[3] A. Bruederle and R. Hodler. Nighttime lights as a proxy for human development at the local level. *PloS one*, 13(9):e0202231, 2018.

[4] X. Chen and W. D. Nordhaus. Using luminosity data as a proxy for economic statistics. *Proceedings of the National Academy of Sciences*, 108(21):8589–8594, 2011.

[5] M. Fatehkia, B. Coles, F. Ofli, and I. Weber. The relative value of facebook advertising data for poverty mapping. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 934–938, 2020.

[6] J. V. Henderson, A. Storeygard, and D. N. Weil. Measuring economic growth from outer space. *American Economic Review*, 102(2):994–1028, 2012.

[7] N. Jean, M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794, 2016.

[8] M. Jerven. *Poor numbers: how we are misled by African development statistics and what to do about it*. Cornell University Press, 2013.

[9] L. Kondmann and X. X. Zhu. Measuring changes in poverty with deep learning and satellite imagery. In *ICLR Practical ML for Developing Countries Workshop*, 2020.

[10] C. Njuguna and P. McSharry. Constructing spatiotemporal poverty indices from big data. *Journal of Business Research*, 70:318–327, 2017.

[11] N. N. Patel, F. R. Stevens, Z. Huang, A. E. Gaughan, I. Elyazar, and A. J. Tatem. Improving large area population mapping using geotweet densities. *Transactions in GIS*, 21(2):317–331, 2017.

[12] A. Perez, C. Yeh, G. Azzari, M. Burke, D. Lobell, and S. Ermon. Poverty prediction with public landsat 7 satellite imagery and machine learning. *arXiv preprint arXiv:1711.03654*, 2017.

[13] N. Pokhriyal and D. C. Jacques. Combining disparate data sources for improved poverty prediction and mapping. *Proceedings of the National Academy of Sciences*, 114(46):9783–9792, 2017.

[14] E. Sheehan, C. Meng, M. Tan, B. Uzkent, N. Jean, M. Burke, D. Lobell, and S. Ermon. Predicting economic development using geolocated wikipedia articles. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2698–2706, 2019.

[15] J. E. Steele, P. R. Sundsøy, C. Pezzulo, V. A. Alegana, T. J. Bird, J. Blumenstock, J. Bjelland, K. Engø-Monsen, Y.-A. de Montjoye, A. M. Iqbal, et al. Mapping poverty using mobile phone and satellite data. *Journal of The Royal Society Interface*, 14(127):1–10, 2017.

[16] H. Taubenböck, J. Staab, X. X. Zhu, C. Geiß, S. Dech, and M. Wurm. Are the poor digitally left behind? indications of urban divides based on remote sensing and twitter data. *ISPRS International Journal of Geo-Information*, 7(8):304, 2018.

[17] I. Tingzon, A. Orden, S. Sy, V. Sekara, I. Weber, M. Fatehkia, M. G. Herranz, and D. Kim. Mapping poverty in the philippines using machine learning, satellite imagery, and crowd-sourced geospatial information. In *AI for Social Good ICML 2019 Workshop*, 2019.

[18] N. B. Weidmann and S. Schutte. Using night light emissions for the prediction of local wealth. *Journal of Peace Research*, 54(2):125–140, 2017.

[19] C. Yeh, A. Perez, A. Driscoll, G. Azzari, Z. Tang, D. Lobell, S. Ermon, and M. Burke. Using publicly available satellite imagery and deep learning to understand economic well-being in africa. *Nature communications*, 11(1):1–11, 2020.

[20] X. Zhao, B. Yu, Y. Liu, Z. Chen, Q. Li, C. Wang, and J. Wu. Estimation of poverty using random forest regression with multi-source data: A case study in Bangladesh. *Remote Sensing*, 11(4):375–393, 2019.